

RESEARCH ON SPOKEN LANGUAGE PROCESSING

Progress Report No. 23 (1999)

Indiana University

**Audio-Visual Perception of Sinewave Speech
in an Adult Cochlear Implant User: A Case Study¹**

Winston D. Goh,² David B. Pisoni,³ Karen I. Kirk,³ and Robert E. Remez⁴

*Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405*

¹ This research was supported by NIH-NIDCD Research Grant DC00111 to Indiana University, NIDCD K23 Research Grant DC00126 to Indiana University School of Medicine, and NIDCD Research Grant DC00308 to Barnard College. We would like to thank Stacey Yount for gathering information on the CI patients' test scores, and Lorin Lachs for helpful comments during the preparation of this article.

² Also, Department of Social Work & Psychology, National University of Singapore.

³ Also DeVault Otologic Research Laboratory, Department of Otolaryngology-Head & Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

⁴ Department of Psychology, Barnard College, New York, NY.

Audio-Visual Perception of Sinewave Speech in an Adult Cochlear Implant User: A Case Study

Abstract. We investigated a post-lingually deafened cochlear implant user's ability to perceive sinewave replicas of spoken sentences. The patient, Mr. S, transcribed sinewave sentences under audio-only (AO), visual-only (VO), and audio-visual (A+V) conditions. His performance was compared to the data collected from a group of normal-hearing participants in an earlier study by R.E. Remez, J.M. Fellowes, D.B. Pisoni, W.D. Goh, and P.E. Rubin (1998). Results showed that Mr. S derived a larger gain from additional visual information provided by the talker's face than the normal-hearing controls. The increase in performance under A+V presentation reflected the superior lip-reading skills that this patient displayed and his ability to use this skill to integrate the information provided by the talking face and the sinewave speech to perceive the underlying sentence. Implications of these findings for multimodal phonetic coherence in speech perception are discussed.

It has long been known that combining audio and visual information facilitates the perception of speech. In their pioneering study, Sumbly and Pollack (1954) demonstrated that the intelligibility of spoken words can be enhanced by as much as +15 dB in noisy environments if listeners are able to see the talker's face. This is a substantial gain in performance that surpasses even the best hearing aid devices. Using visual information from a dynamically articulating face for phonetic and lexical identification is a skill which almost everyone will benefit from when listening in noisy environments, especially when people get older and their hearing deteriorates (Summerfield, 1987). For the hearing-impaired population, the visual route may play an especially major role in speech perception. Some of the speech cues for consonants that are difficult to hear are easy to see and vice-versa (Walden, Prosek, Montgomery, Scherr, & Jones, 1975). For example, /f/ and /θ/ are auditorily confusable, but they are very distinct visually when the talker's articulatory movements can be seen. The enormous gain from seeing the talker's face is eloquently captured by a question frequently asked of hearing-aid practitioners – "Doctor, why can I understand you so much more clearly when I wear my glasses?" (Summerfield, 1987). Research into the nature of audio-visual integration in speech perception can therefore provide substantial insights and applications for rehabilitative procedures, techniques, and training methods to assist the hearing impaired. The study of multimodal speech perception also raises many important theoretical issues about the scope and domain of current models of speech perception and spoken language processing (see Berstein, Demorest, & Tucker, in press; Massaro, 1998).

The absolute gain in performance observed from the visual aspects of speech is highest in a noisy environment or in other conditions that make auditory perception difficult (Sumbly & Pollack, 1954). Therefore, the best way to observe the influence of visual information is to look at identification performance with impoverished auditory stimuli. The traditional way of studying this problem in the past was to manipulate the signal-to-noise ratio (SNR) for the environment in which the speech stimuli is presented. Another way involved reducing the amount of information that is normally available in the auditory speech waveform. One such technique is to use sinewave speech instead of natural speech (Remez, Rubin, Berns, Pardo, & Lang, 1994; Remez, Rubin, Pisoni, & Carrell, 1981).

In sinewave speech, time-varying sinusoidal waveforms are generated by a digital synthesizer to match the LPC-derived center frequencies and amplitudes of the formants in the natural utterance. The synthetic sinewave pattern preserves the dynamics of frequency and amplitude variations observed in natural speech over time, but differs from natural speech in several important ways. There are no

harmonics, broadband formant structures, formant frequency transitions, steady-state formants, or changes in fundamental frequency. In short, sinewave speech patterns contain none of the “traditional” speech cues that are assumed to form the basis of speech perception – e.g., formant frequency transitions that cue manner and place of articulation (see Remez et al. 1981).

Despite the unnatural characteristics of sinewave speech, these sound patterns are still intelligible (Remez et al., 1981; 1994). The absence of traditional acoustic cues for phonetic perception implies that sinusoidal replicas of speech should be perceived as independently changing tones and not as an integrated, linguistic percept. However, listeners are still able to extract the phonetic and lexical properties of the utterance from the highly impoverished, skeletal representation of the natural token that is preserved in the sinewave replica. This result suggests that sufficient phonetic information is still encoded in the relational and time-varying structure that is represented in the sinewave pattern, even though the synthetic waveform is obviously not producible by a vocal tract. Sinewave speech perception also shows the multimodal facilitation observed for natural speech (Remez, Fellowes, Pisoni, Goh, & Rubin, 1998). A considerable increase in identification performance was found when the sinewave patterns are presented in an audio-visual context compared to an audio-only context.

Previous studies on sinewave speech perception have so far used only participants who have normal hearing at the time of testing. Since audio-visual speech perception may be even more critical for people who have hearing impairment, it is important to begin investigations into how members of this clinical population perceive sinewave speech. In particular, how would hearing-impaired individuals fitted with a cochlear implant (CI) fare in listening to sinewave speech under different presentation conditions? We are especially interested in patients who perform very well with their CI and who demonstrate the ability to use visual information to mitigate their hearing impairment. Would such users be able to integrate visual information with very unnatural auditory patterns? In this paper, we report the performance of one patient, Mr. S, in transcribing sinewave sentences under audio-only (AO), visual-only (VO), and audio-visual (A+V) conditions and then compare his performance to a group of normal-hearing participants whose data was collected by Remez et al. (1998).

Patient Background

Our patient, Mr. S, is a 35-year-old Caucasian male with a graduate degree. He has a profound hearing loss due to cryoglobulinemia and autoimmune syndrome. Onset of his deafness occurred in 1993 when he was 29. His hearing impairment was diagnosed as a profound loss a year later and he was implanted with a Clarion 8-channel CI in 1995. He has been using the CI for the past 4 years and is considered to be an exceptionally good user by the clinical staff. We will now describe Mr. S’s performance on a battery of standard clinical tests that were collected in 1998. All tests were conducted while he was using the CI.

The Iowa Consonant Test (Tyler, Preece, & Tye-Murray, 1983) is a closed-set test of consonant recognition in which the listener is familiarized with 16 different consonants in the /aCa/ format. The listener is then asked to identify the consonant he hears out of a choice of 16 alternatives. Chance performance for consonant identification is approximately 6%. This test can also be analyzed in terms of the listener’s ability to identify phonetic features. Chance performance for consonant voicing, manner, and place of articulation identification is 50%, 33%, and 20% respectively. On the Iowa Consonant Test, Mr. S achieved a total score of 79% correct, 96% on voicing, 94% on manner, and 85% on place.

The CUNY Sentences Test (Boothroyd, Hannin, & Hnath, 1985) is an open-set sentence recognition task in which the listener is presented with sentences in three listening conditions: AO, VO, and A+V. The test is scored in terms of the total number of words correctly identified. On this test, Mr. S

obtained a perfect score of 100% in the A+V condition, 92% in the AO condition, and 63% in the VO condition.

Table 1 compares Mr. S's scores on these tests and the average scores of 28 other CI patients. Table 1 also lists the performance of Mr. S and the other CI patients in a recent study (Kaiser, Kirk, Pisoni, & Lachs, 2000) that tested the participants' ability to identify isolated consonant-vowel-consonant (CVC) words from the Hoosier audiovisual multi-talker database (Lachs & Hernandez, 1998; Sheffert, Lachs, & Hernandez, 1997) under AO, VO, and A+V presentations, using both single-talker and multiple-talker presentation conditions. Generally, Mr. S's performance on these speech perception tests indicate that he is able to perceive speech without any content cues in a controlled test environment. It is clear that Mr. S is an exceptionally good implant user relative to the other CI patients. His lip-reading performance is consistently at least two standard deviations higher than the average CI patient, as shown in his scores for the various tests in the VO conditions in Table 1.

	Mr. S	Other CI Patients ($N = 28$)	
		<i>M</i>	<i>SD</i>
Iowa Consonant Test*			
Total	79	45.0	17.3
Voicing	96	88.7	12.9
Manner	94	67.1	15.7
Place	85	52.3	16.3
CUNY Sentences Test**			
Audio-only (AO)	92	55.0	29.6
Visual-only (VO)	63	24.3	15.6
Audio-visual (A+V)	100	91.2	9.1
Indiana Multi-talker Isolated CVC Words***			
Single talker condition			
Audio-only (AO)	67.0	30.1	19.1
Visual-only (VO)	30.5	15.7	5.1
Audio-visual (A+V)	88.5	69.0	13.8
Multiple talker condition			
Audio-only (AO)	55.5	30.00	16.57
Visual-only (VO)	30.5	14.79	7.07
Audio-visual (A+V)	86.0	60.03	14.60

Table 1. Comparison of Mr. S's percent correct scores and the average percent correct scores of other CI patients' on several speech perception tests.

* Tyler, Preece, and Tye-Murray (1983).

** Boothroyd, Hannin, and Hnath (1985).

*** From Kaiser, Kirk, Pisoni, and Lachs (2000).

Mr. S also achieved an auditory digit-span score of 10 on the WISC-forward and 8 on the WISC-backward collected under AO conditions. On the word familiarity test (FAM; Lewellen, Goldinger, Pisoni, & Greene, 1993), which indexes subjective familiarity with words of varying frequencies, he had a mean FAM score of 3.57 on low frequency words, 6.39 on mid-frequency words, and 6.85 on high

frequency words. These scores are comparable to the average scores obtained for normal-hearing, high-vocabulary participants as described in Lewellen et al. (1993). His performance on the WISC digit-span and FAM tests indicate that his short-term memory capacity and word familiarity are comparable to normal-hearing subjects.

Method

Participants

The normal-hearing participants whose data we used as a comparison group consisted of 25 young adults from the Indiana University community. These participants were a subset of the sample that participated in the study described in Remez et al. (1998). All participants were native speakers of English and reported normal hearing and normal vision or corrected-to-normal vision at the time of testing. None of the participants had any previous exposure or familiarity with sinewave analogs of speech signals. All participants were students enrolled in Introductory Psychology classes. They received either course credit for participation or they were paid as a volunteer. Our patient, Mr. S, was paid as a volunteer. He also had no prior experience with sinewave speech before the present tests and he had corrected-to-normal vision. He was tested on the sinewave speech in August 1999.

Apparatus and Materials

The 18 sentences used in the present study were obtained from the database developed by Remez et al. (1998) and are listed in the Appendix. The original sentences were recorded and digitized and then an expert phonetician analyzed the sampled data to estimate the formant center frequencies and amplitudes. Formant center frequencies were obtained by comparing discrete Fourier spectra and linear prediction estimates. The synthesis parameters were created by tracing the formant patterns over time. The fundamental frequency of phonation was estimated from a narrow-band Fourier representation of the natural spectra. A software synthesizer was then used to convert the frequency and amplitude values taken at 10 msec intervals for F0, F1, F2, F3 and fricative formants to time-varying sinusoids (Rubin, 1980). The sinewave replicas of each sentence were composed of tone analogs of the three oral formants. A fourth tone was used to reproduce fricative formants when these were present and discontinuous with the oral formants.

The first 8 sentences were spoken by one of the authors (RER), and were used for the familiarization phase and AO condition. The other 10 sentences were spoken by an adult female speaker whose natural speech intelligibility had been verified by other normal-hearing volunteer participants as acoustically intelligible (see Bradlow, Torretta, & Pisoni, 1996). The female speaker's sentences were used in the VO and A+V conditions. The sinewave patterns for her sentences were combined and synchronized with the video clips using Adobe Premiere 4.2. All stimulus materials were presented to participants via a Macintosh Quadra 950 machine with a Targa 2000 video card. This system presented the 14-bit color video samples at 30 frames per second in full-screen mode at 640x480 resolution on a 17-inch monitor.

Design and Procedure

The 18-sentence presentation sequence was fixed for all participants. All normal-hearing participants listened to the audio track via a pair of Beyer Dynamic DT100 headphones. The audio stimuli were presented at approximately 75 dB SPL. Our patient, Mr. S, listened to the audio track via a set of Labtec LS-1020 desktop computer speakers with his CI turned on. Prior to the start of the experiment, we

calibrated the output amplitude of the speakers to a signal level where he could correctly identify five auditorily presented, naturally spoken CVC words in a row.

The sequence of testing for Mr. S was as follows. The first three sentences were used as a familiarization sequence to acclimatize him to the unnatural timbre of the sinewave sentences. These materials were presented audio-only and the sentences were already transcribed for him on the answer sheet. Each sentence was repeated five times with 10 seconds between repetitions and 20 seconds between sentence blocks. A warning tone occurred before the start of a new sentence block. The next five sentences followed the same procedure and comprised the AO condition. Our patient was asked to transcribe these sentences while he listened. For the familiarization phase and the AO condition, the video monitor remained blank while the audio signals were played out via the speakers. The control participants in Remez et al. (1998) followed precisely this same procedure except that the signals were presented over individual headphones and the participants were run in groups of five or smaller.

In Remez et al. (1998), the VO and A+V conditions were run between-subjects using all ten of the female talker's sentences. It was obviously not possible to follow this same procedure with our patient so several changes were made in the presentation format. For our CI patient, after the AO condition, the first five sentences of the female talker were presented in the VO condition, followed by the last five sentences of that same talker in the A+V condition. The CI patient and the normal-hearing control participants were instructed to look at the video monitor and write down their responses only during the intervals between repetitions and sentence blocks.

Scoring

The number of syllables correctly transcribed was used as the dependent measure for both the CI patient and the normal-hearing control group. Because there were some small procedural differences between the Remez et al. (1998) data collection and the session with our CI patient, we had to ensure that the control data obtained from the earlier study was a valid comparison to make. For the control participants who were assigned to the VO condition in Remez et al., we only scored their responses for the *first* five sentences, since these were exactly the same sentences presented to the CI patient in the VO condition. Conversely, for the control participants assigned to the A+V condition in Remez et al., we only scored their responses for the *last* five sentences, since these were the same sentences used in the CI patient's A+V condition. All five sentences were scored in the AO transcription of RER's sentences because these sentences were identical for both the CI patient and the participants in Remez et al. These scoring procedures ensured that appropriate comparisons could be made between our patient's responses and those obtained from the normal-hearing participants.

Results

The transcription performance of our CI patient, Mr. S, and the relevant data from the normal-hearing participants from Remez et al. (1998) are summarized in Table 2. The results show that the CI patient's AO performance (53%) was not as good as the average of the normal-hearing controls (65%). It should be emphasized here that the CI patient's performance is, however, within one standard deviation of the mean of the normal-hearing controls' performance. It is very likely that the inherent difficulty in perceiving sinewave speech is further compounded by reliance on a CI device for speech perception. Normal-hearing participants would not have as much difficulty because of their intact hearing abilities. The average performance of the normal-hearing controls displayed here is very similar to the previous results reported for AO listening conditions (Remez et al., 1981).

	Audio-only (AO)	Visual-only (VO)	Audio-visual (A+V)
Mr. S	52.5	43.2	89.7
Normal-hearing*			
<i>M</i>	64.7	18.0	85.6
<i>SD</i>	24.4	10.6	13.5
<i>N</i>	25	14	11

Table 2. Average percentage of correct syllables for Mr. S and normal-hearing listeners for the three presentation conditions.

* from Remez, Fellowes, Pisoni, Goh, and Rubin (1998).

For the VO condition, it is clear that Mr. S showed superior performance (43%) relative to the normal-hearing controls (18%). His performance is more than two standard deviations from the mean of the normal-hearing listeners. This is probably due to the enhanced lip-reading abilities that are typical of the hearing-impaired population (Summerfield, 1987). Of greater interest to us, however, is his performance in the A+V condition. With the addition of visual information, our CI patient's performance (90%) is slightly above the average performance of the normal-hearing controls (86%). This patient is clearly able to use and integrate information from the dynamically changing articulators in the video display with the auditory information, despite the unspeechlike qualities of the auditory sinewave speech patterns. The additional visual information drives his performance up to levels that are comparable to the average performance of a group of normal-hearing controls.

One of the more interesting questions about Mr. S's performance deals with assessing the contribution of the visual information relative to the possible information available in the absence of visual stimulation (see for example Sumbly & Pollack, 1954). The actual contribution from the visual modality is the difference in scores between the A+V and AO conditions. The possible available information is the difference between the total information possible and the performance in the AO conditions, i.e. 100% minus the AO performance. The ratio of the actual contribution to the possible available information is the amount of gain obtained from seeing the talker's face, and can be considered a measure of visual enhancement. This ratio normalizes for absolute differences in AO performance and is formalized below:

$$R = (A+V - AO) / (100 - AO)$$

Using this metric, our patient displays a 78.3% gain from seeing the talker's face, compared to a 59.2% gain for the mean of the normal-hearing controls. This difference in visual enhancement suggests that Mr. S is deriving a much larger benefit from the visual information than the normal-hearing controls to achieve his exceptionally high level of performance in the A+V condition. This level of performance is clearly not due to his superior lip-reading ability alone, since he was only performing at 43% accuracy in the VO condition, although this ability is probably an influential factor. The improvement observed under A+V presentation is due to his ability to use and integrate the information provided through the visual modality with the cues provided by the time-varying dynamics of the tone analogs to perceive the underlying sentences.

Discussion

If sinewave replicas can be perceived as speech even though a natural source cannot be attributed to the signal, the results imply that the perceptual organization of speech depends on the establishment of coherence among dissimilar sensory elements (Remez et al., 1994) – in the case of sinewave speech, it is

the coherence of independently changing tones. In multimodal contexts, the visual perception of a dynamically articulating face combined with the auditory perception of the very unspeechlike quality of a sinewave signal makes the event especially incoherent. In other words, no natural talker can be made to appear to an observer as the original source of the sinewave speech. Thus, multimodal presentation provides a strong test of phonetic coherence despite the incoherence of the visual and auditory information as an integrated perceptual event. In Remez et al. (1998), the combination of the video signal plus an auditory track containing a single tone analog of the F2 formant was the only single-tone condition that showed a facilitation over the VO condition. The interpretation of this result by Remez et al. was that both the F2 sinewave analog and the visual information provided congruent cues regarding the underlying gestures for place of articulation, thus producing multimodal phonetic coherence of the perceptual event from two separate and independent sensory inputs.

In this study, multimodal perception of sinewave speech was investigated for the first time in a CI user. Although Mr. S did not perform as well as the normal-hearing controls in the AO condition, his performance was comparable or slightly better than the normal-hearing controls in the A+V condition. Clearly, the additional complementary visual information in the latter condition allowed both Mr. S and normal-hearing participants to better perceive the sinewave sentences. More importantly, however, Mr. S obtained a much larger benefit from seeing the talker's face than the normal-hearing controls (78% gain versus 59% gain, respectively). This pattern is consistent with previous findings showing that the visual modality plays a large role in the speech perception of the hearing-impaired population (Summerfield, 1987). It is also important to emphasize here that the task used in the present study was to transcribe highly impoverished sinewave speech patterns, not normal or natural speech samples. Mr. S's ability to simultaneously integrate visual information with the sinewave tracks of the auditory signal is impressive and suggests that multimodal phonetic coherence can occur even when a profoundly hearing-impaired listener perceives highly impoverished sound patterns through a CI device. These findings from Mr. S, a patient with a CI, provide additional support for the original proposal of Remez et al. (1981) – speech perception can occur without traditional speech cues. Elements of speech perception from multiple sensory modalities depend on the establishment and preservation of perceptual coherence among individual elements and attributes at a more abstract level that reflects the underlying common source of the speech signal, as represented in the talker's articulatory gestures (Remez et al., 1994; 1998).

We are confident that this exploratory case study will spur future research on sinewave speech perception in the hearing-impaired and other clinical populations such as the elderly and language-delayed children. The use of sinewave speech patterns with these populations will provide useful converging evidence about the similarities and differences in the nature of speech perception processes in different populations who have known and well-documented sensory, perceptual and cognitive impairments in the ability to encode and perceive sensory input. Data from these listeners should therefore provide valuable new insights into the multimodal organization of speech and spoken language processing.

References

- Bernstein, L.E., Demorest, M.E., & Tucker, P.E. (in press). Speech perception without hearing. *Perception & Psychophysics*.
- Boothroyd, A., Hannin, L., & Hnath, T. (1985). A sentence test of speech perception: Reliability set equivalence and short term learning (internal report RCI 10). New York: City University of New York.

- Bradlow, A.B., Torretta, G.M., & Pisoni, D.B. (1996). Intelligibility of normal speech I: Global and fine grained acoustic-phonetic talker characteristics. *Speech Communication, 20*, 255-272.
- Kaiser, A., Kirk, K.I., Pisoni, D.B., & Lachs, L. (2000). Audiovisual speech integration in adults with cochlear implants or normal hearing: Lexical and talker effects. Paper to be presented at the ARO midwinter meeting, St. Petersburg Beach FL, February 20-24 2000.
- Lachs, L., & Hernández, L. R. (1998). Update: The Hoosier audiovisual multi-talker database. In *Research on Spoken Language Processing Progress Report No. 22* (pp. 377-388). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Lewellen, M.J., Goldinger, S.D., Pisoni, D.B., & Greene, B.G. (1993). Lexical familiarity and processing efficiency: Individual differences in naming, lexical decision, and semantic categorization. *Journal of Experimental Psychology: General, 122*, 316-330.
- Massaro, D.W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge: MIT Press.
- Remez, R.E., Fellowes, J.M., Pisoni, D.B., Goh, W.D., & Rubin, P.E. (1998). Multimodal perceptual organization of speech: Evidence from tone analogs of spoken utterances. *Speech Communication, 26*, 65-73.
- Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S., & Lang, J.M. (1994). On the perceptual organization of speech. *Psychological Review, 101*, 129-156.
- Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech perception without traditional speech cues. *Science, 212*, 947-950.
- Rubin, P.E. (1980). Sinewave synthesis. Internal memorandum. New Haven: Haskins Laboratories.
- Sheffert, S., Lachs, L. & Hernandez, L. R. (1997). The Hoosier audiovisual multi-talker database. In *Research on Spoken Language Processing Progress Report No. 21* (pp. 578-583). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212-215.
- Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3-51). Hillsdale: Erlbaum.
- Tyler, R.S., Preece, J., & Tye-Murray, N. (1983). *The Iowa cochlear implant tests*. Iowa City: Department of Otolaryngology - Head and Neck Surgery, University of Iowa.
- Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., & Jones, C.J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research, 20*, 130-145.

Appendix

Sentence Materials (adapted from Remez et al., 1998)

Familiarization

The bill was paid every third week.
The soft cushion broke the man's fall.
Two blue fish swam in the tank.

Audio-only (AO)

A small creek cut across the field.
The fruit peel was cut in six slices.
Her purse was filled with useless trash.
The stray cat gave birth to kittens.
Where were you a year ago?

Visual-only (VO)

Always close the barn door tight.
This is a grand season for hikes on the road.
He ran halfway to the hardware store.
Kick the ball straight and follow through.
The term ended in late June that year.

Audio-visual (A+V)

Use a pencil to write the first draft.
Cut the pie into large parts.
The boy was there when the sun rose.
A cup of sugar makes sweet fudge.
What joy there is in living.